Solar Power Generation Model Report

Billie Kim

1 Introduction

Solar plants convert sunlight into renewable electricity by harvesting solar energy through the use of solar panels. These solar panels are made up of solar modules which are assemblies of connected photovoltaic (PV) solar cells. These types of systems can generate clean, renewable electricity that can power homes and businesses in a cost-effective and efficient way. However, common problematic issues such as inverter failure, sensor issues, temperature variability, and maintenance of dirty solar panels can cause misproduction which can critically disrupt the performance of an entire grid system. Therefore, solar generation forecasting is crucial for planning and maintaining an efficient grid. We use data collected from two solar power plants in India over a 34 day period— from May 15, 2020 to June 17, 2020— to create a model to predict solar power generation. Each plant has a power generation dataset at the inverter level as well as a weather sensor dataset at a plant level.

2 Data Wrangling

We imported four CSV files from kaggle— two power generation datasets and two weather datasets[1]. The power generation datasets contain information from 22 inverters at every 15 minute intervals with numerical features such as DC power, AC power, daily yield, and total yield, as well as other metadata

[1] https://www.kaggle.com/datasets/anikannal/solar-power-generation-data

such as plant ID and source key. The weather datasets were also compiled in a similar manner where each row represents each 15 minute increment in time and contains features such as ambient temperature, module temperature, and solar irradiation. Below are descriptions and units of measurement for each variable within the data that we acquired:

- **DC Power** (kW): The amount of direct current energy generated by the inverter
- **AC Power** (kW): The amount of alternating current generated by the inverter
- **Daily Yield** (kW): The cumulative sum of power generated on that day, till that point in time
- **Total Yield** (kW): The total yield for the inverter till that point in time
- **Ambient Temperature** (°C): The ambient temperature at the location of the plant
- **Module Temperature** (°C): The temperature sensor reading of the solar panel module
- **Irradiation** (W/m^2): The amount of light intensity from the sun

While the data was relatively clean to begin with, we performed some initial cleaning prior to moving forward such as checking for duplicates, dropping missing values, and converting the datetime column into a pandas datetime index. For each power plant, we merged the power and weather datasets into a single dataframe and then created plant-wide average dataframes to work with going forward.

3 Exploratory Data Analysis

In a typical solar generating system, solar energy is converted into DC power by solar panels and is then converted into AC power by inverters. We first visualized the relationship between DC power and AC power and noticed that DC power was 10 times larger than AC power in plant 1. The calculated inverter conversion efficiency for plant 1 was 9.78% and 97.8% for plant 2. This finding raised some alarms since the conversion efficiency for solar inverters are typically around 97% to 99%— meaning energy loss is relatively minor. In other words, it is either that plant 1 experiences substantial loss from DC-AC conversion due to faulty equipment or caused by human error during the data collection phase. We determined the best way to move forward was to rescale plant 1's DC power with the conversion efficiency ratio using simple statistics (Figure 1). By doing so, we see a much more normal relationship between AC and DC power for plant 1.

We also found that when exploring AC and DC power during day hours, we see a nice bell-shaped curve in plant 1; however, plant 2's AC and DC power seemed to be capped at a threshold (Figure 2). We wanted to examine the root cause of this event, so we examined how each inverter performed during day hours. In doing so, we found that plant1's DC power generation per inverters appear to be stable and tightly grouped with slightly lower signals from two inverters: "1BY6WEcLGh8j5v7" and "bvBOhCH3iADSZry" (Figure 3). However, we discovered that there is much more variation in plant 2's inverters which could be an indication that plant 2's inverters need to be cleaned and scheduled for maintenance. We noticed substantially lower signals from four underperforming inverters: "Et9kgGMDl729KT4", "LYwnQax7tkwH5Cb", "Quc1TzYxW2pYoWX", and "rrq4fwE8jgrTyWY".

4 Target Feature Transformations, Outlier Detection, and Correlations

Daily yield is the cumulative sum of power generated on that day and at that point in time— meaning that daily yield resets at the end of each day. In its raw state, it wasn't quite useful or suitable for

modeling. Therefore, we transformed our target variable into 'delta daily yield' to find the change in power generated within each 15 minute increment in time.

After visualizing 'delta daily yield', we found large spikes as potential outliers. After calculating our target feature's mean and standard deviation, we replaced outliers that were three standard deviations above the mean with its median value (Figure 4).

We created a feature correlation matrix to identify patterns and to see the relationships between each variable (Figure 5). The features that seem to be most important to our target variable 'delta_daily_yield2' are 'dc_power', 'irradiation', 'module_temperature', and 'ambient_temperature'. We can clearly see how applying appropiate transformations to 'daily_yield' and removing outliers increased the correlations among the features.

5 Pre-Processing and Training

We split our training and testing data. The last seven days from June 11, 2020 to June 17, 2020 are held as our test set. We used the mean of 77.82 as a baseline performance indicator and calculated a mean absolute error (MAE) of 79.8 which tells us that if we use the mean as the best guess for the change in daily yield for every 15 minutes we would expect to be off by around 79.8 kilowatts. Going forward, we will use machine learning algorithms to decrease the loss function for our model and to minimize prediction errors.

6 Modeling

Three models were compared: Linear Regression, Random Forest, and Gradient Boosting. We created pipelines for each model where we scale our features and tune hyperparameters. We found consistent r-squared values of over 98% in all models. When we looked at feature importances of each model, we found DC power and module temperature to be the most important features. We used 5-fold cross-validation to compare test performance across each of the models (Figure 6). We found a CV mean MAE score of 6.31 for the Linear Model and 4.94 for the Random Forest Model. The Gradient Boosting Model produced the lowest MAE score of 4.86.

We were able to find new weather data for the next seven days from June 18, 2020 to June 24, 2020[2] . However, the data contained only ambient temperature and irradiation variables. Therefore, we created simple linear models to find inputs for module temperature and dc power to feed into our solar power model to predict delta daily yield (Figure 7). We aggregated and summed our 'delta daily yield' predictions by date to get the cumulative daily yield predictions for the next 7 days (Figure 8).

7 Conclusion

We showed how our model could use forecasted weather data to predict solar power production of a solar power plant. Plant 1's data was relatively clean and easier to work with and model; however, plant 2's data contained far greater amounts of outliers and extreme data points. In the future, we hope that we can further work on this project and explore different ways to tackle the data quality issues that we found in plant 2. For example, implementing a stricter rule or plan for handling outliers. We also would like to explore other weather parameters as inputs to our model such as relative humidity and wind speed which are also factors that affect solar power generation.

[2]https://weatherdownloader.oikolab.com/downloader

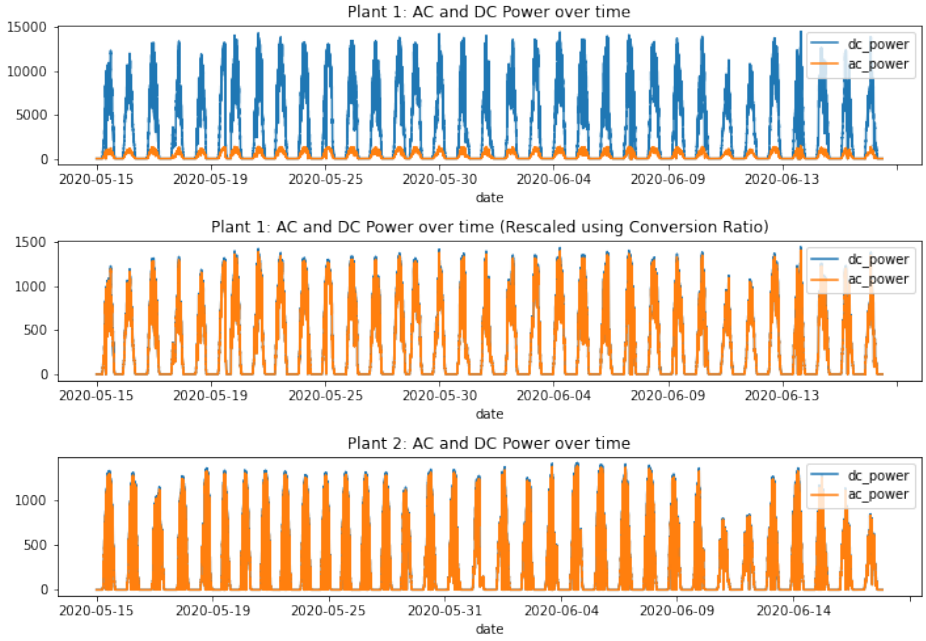

FIGURE 1: Plant 1 & 2: AC and DC Power over time

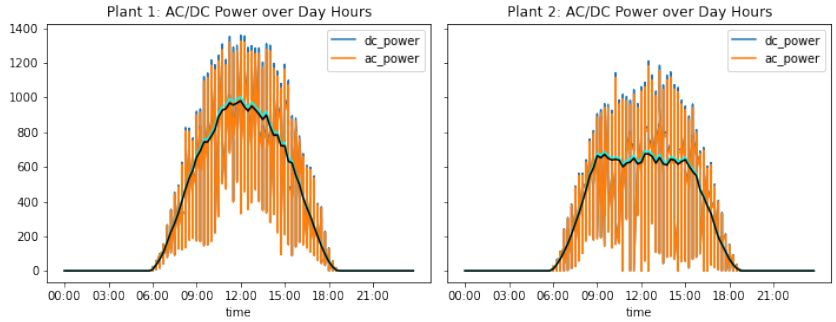

FIGURE 2: Plant 1 & 2: AC/DC Power over day hours

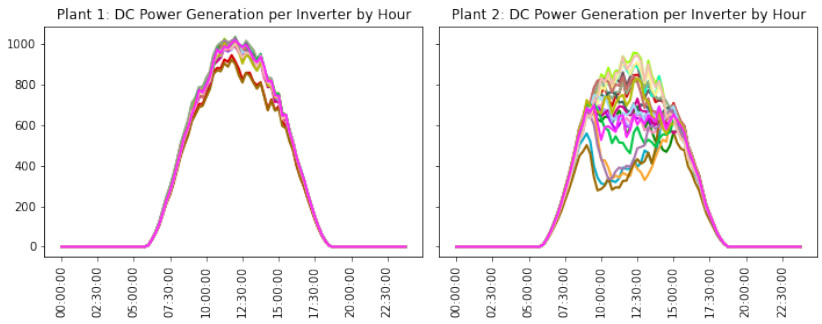

FIGURE 3: Plant 1 & 2: DC Power Generation per Inverter by hour

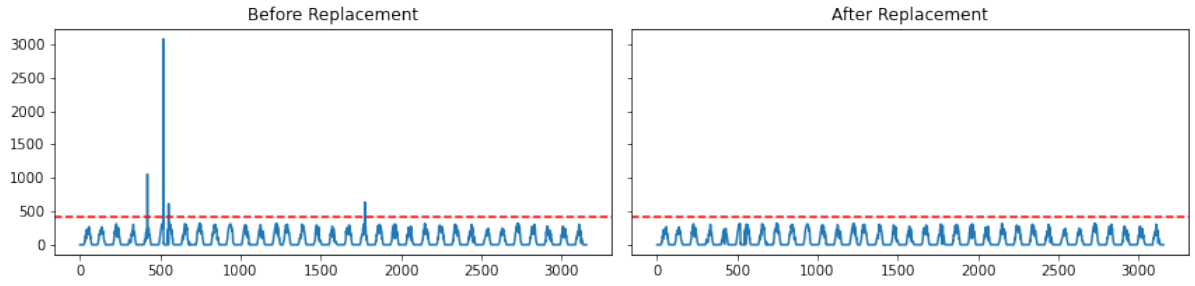

FIGURE 4: Plant 1: Outlier Detection and Replacement

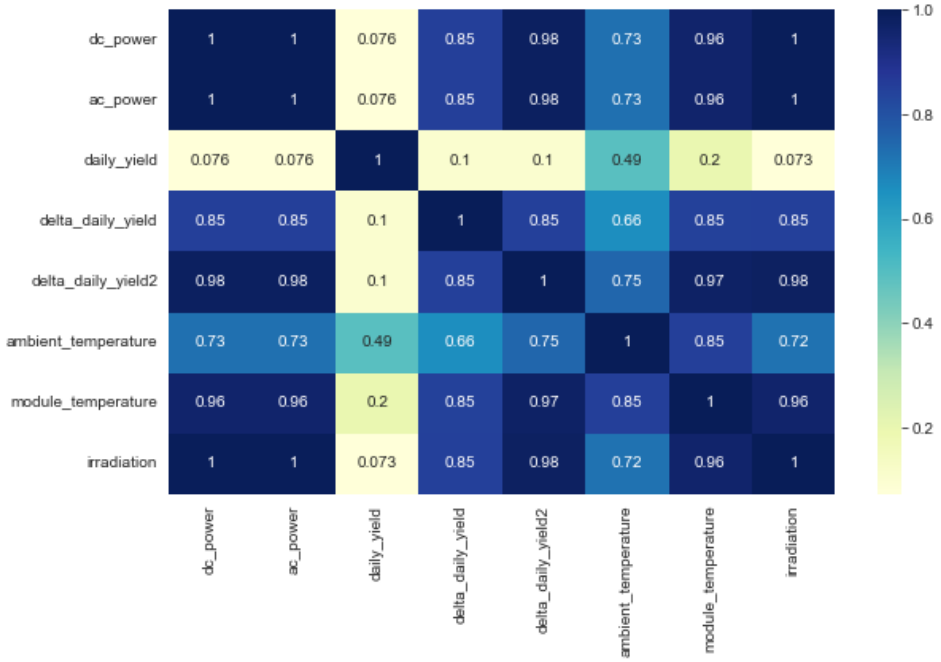

FIGURE 5: Plant 1: Feature Correlation Heatmap

	r_squared		Mean Absolute Error		Mean Standard Error	
Plant 1 Models	**r2 (mean)**	**r2 (std)**	**MAE (mean)**	**MAE (std)**	**MSE (mean)**	**MSE (std)**
Linear Regression	0.9878	0.0019	6.3149	1.4682	90.1842	26.7941
Random Forest	0.9883	0.0025	4.9350	1.7925	88.5963	31.8255
Gradient Boosting	0.9885	0.0033	4.8615	1.8495	86.4643	34.9131

FIGURE 6: Cross-Validation Metrics Results on Test Data

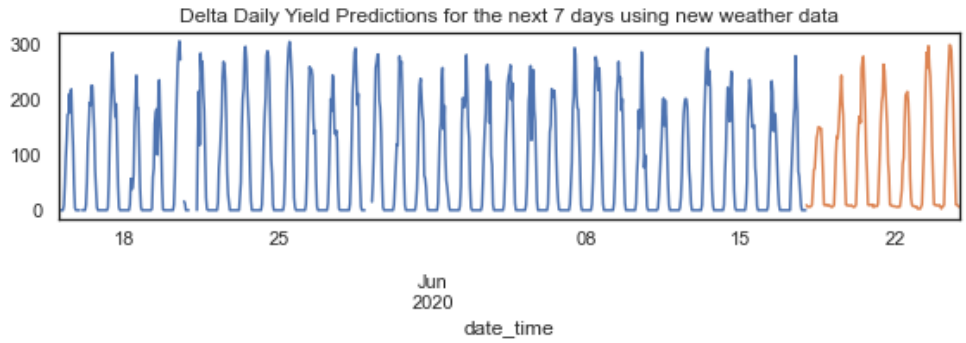

FIGURE 7: Visualizing 'delta daily yield' Predictions for the next 7 days

date	predicted_daily_yield
2020-06-18	1435.649774
2020-06-19	1861.912548
2020-06-20	2175.264188
2020-06-21	2052.727623
2020-06-22	1546.800547
2020-06-23	2395.338236
2020-06-24	2477.258647

FIGURE 8: Cumulative Daily Yield Predictions for the next 7 days